A Methodology for Encoding Problem Lists with SNOMED CT in General Practice

Francis Lau, Ph.D., Ray Simkus, M.D., Dennis Lee, M.Sc.
School of Health Information Science, University of Victoria, Victoria, B.C., Canada
fylau@uvic.ca, ray@wmt.ca, dlhk@uvic.ca

ABSTRACT

This paper describes a methodology for encoding problem lists used in general practice with SNOMED CT. Our intent is to help general practitioners to incorporate SNOMED CT into their existing Electronic Medical Record (EMR) systems with minimal disruption as a first step, thus allowing them to assess its impact prior to full-scale conversion. We started with 1,713 original unique terms that made up the problem lists from the general practice EMR used in the study. We ended with 1,468 unique concepts after two cycles of matching and revisions that led to 1,347 or ~92% successful matches. The remaining terms were revised to tease out modifiers or secondary concepts that could be used to provide equivalency through post-coordination. While skeptics of reference terminology systems often balk at their unwieldy size and complexity for local adoption, this study has demonstrated that, using our methodology, it is possible to create a manageable subset of SNOMED concepts for problem lists used in general practice with immediate tangible value.

INTRODUCTION

The problem list is the keystone of the medical record. In general practice settings, the type of problems presented by patients can be quite diverse. Examples range from non-specific symptoms such as headaches with unknown cause, to a diagnosis of coronary disease that can be expressed in different ways such as heart attack and myocardial infarction. The choice of terms used in problem lists becomes an important design issue for the electronic medical record (EMR), since the level of granularity selected for defining the problems and the actual terms entered into the system can affect one's ability to retrieve the information afterwards, thus impacting the overall quality of the EMR system.

There have been many studies on the design and use of controlled terminology to encode the problem lists in EMR systems and their impact on practice [1-8]. Most of these studies are focused on large institutions involving a substantive number of clinical terms in order to accommodate the needs of a wide range of clinicians in the institution. For example in their study of diagnosis and problem lists in a computerized physician order entry system, Wasserman [9] reported that 88.4% of their 8,378 clinical terms were found in SNOMED CT. With the addition of 145 site-specific terms they were able to achieve 98.5% overall content coverage. With the formation of the International Health Terminology Standards Organization (IHTSDO), the historical barriers to SNOMED CT related to cost and the proprietary nature of the product have now been removed, and national initiatives related to EMR's are emerging to use SNOMED CT as a clinical terminology in several countries around the world.

Despite such impressive development, the effort to adopt SNOMED CT in Canada has been minimal to date. There continues to be a concern especially in the primary care setting where most general practices are made up of small groups of practitioners, of whom few are equipped with an EMR. Critics often balk at the enormous size and complexity of SNOMED CT, considering it as too unwieldy and costly for local adoption and use. But a review of data collected from several sites by one author showed the number of codes needed to cover disorders of at least 1:100,000 occurrence would be under 5,000 [10]. Work is underway with IHTSDO and the WICC group of WONCA to finalize this list as a potential primary care SNOMED subset [11].

In this paper, we describe a methodology that we have developed based on an ongoing study to encode problem lists using SNOMED CT (July 2007 release) for a local general practice in Canada. The intent of this methodology is to enable general practitioners to incorporate SNOMED CT into their existing EMR systems within minimal disruption as a first step, thus allowing them to assess its potential impact prior to full-scale conversion.

METHODS

Design and Setting
For this study, we included all the problem list (PL) terms from the commercial EMR system used by a local general practice in British Columbia, Canada. This setting is typical of many general practices

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

across the country, which are made up of small groups of general practitioners working in a private medical office, mostly on a fee-for-service basis. The medical office in this study has four general practitioners who have worked as a group for 30 years in a township with a population of 100,000 located east of Vancouver, British Columbia. The practice has had 8 years of experience using an EMR. At least two of the practitioners record all of the information on their patients on a daily basis at the time of encounter or shortly thereafter. Laboratory and imaging results and consult reports from external sources – both electronic and on paper – are entered into the EMR either by the practitioners themselves or the medical office assistant.

Matching Algorithms
We applied four matching algorithms used in an earlier SNOMED CT to ICD-10 mapping project to find matching SNOMED concepts for each of the PL terms [12]. Three are lexical techniques for exact-match, match-all and partial-match. The fourth is semantic matching that involves retrieving the current concepts based on historical relationships if the initial SNOMED concepts found were inactive. These algorithms are summarized in Table 1.

Algorithm	Explanation
1. Exact match	Exact string match where all words are same and in same sequence, including punctuation
2. Match all	String match where all words are same but not necessary in same order; additional words allowed
3. Partial match	String match where one or more words is found
4. Semantic match	For inactive concepts use historical relationships Was-A, Same-As, May-Be-A, Replaced-By to find current concepts
5. Unmatched	Assigned when no match is found

Table 1. Matching algorithms used in this study

Normalization Steps
In addition to applying the matching algorithms to the original PL terms, we reran the algorithms after we normalized the PL and SNOMED terms to remove "noise" using the Unified Medical Language System (UMLS 2007 version) normalization steps, shown in Table 2a [13,14]. To improve matching, we expanded step-2 to remove both "stop words" and "exclude words" and SNOMED prefixes, shown in Table 2b. For step-5 we included the lookup and stemming methods to uninflect the phrase. The lookup method uses the UMLS SPECIALIST Lexicon's inflection table with ~1 million entries, whereas the stemming method is a computational technique that reduces word variants to a single canonical form [15,16].

No	Step	Example
1	Remove genitive	Hodgkin's disease, NOS → Hodgkin diseases, NOS
2	Remove stop words	Hodgkin diseases, **NOS** → Hodgkin diseases,
3	Convert to lowercase	**H**odgkin diseases, → hodgkin diseases,
4	Strip punctuation	hodgkin diseases**,** → hodgkin diseases
5	Uninflect phrase	hodgkin disease**s** → hodgkin disease
6	Sort words	**hodgkin disease** → disease hodgkin

Table 2a. UMLS normalization steps [8, slide20]

Matching PL Terms
The process of matching the PL terms involved cycling through the matching algorithms one at a time to find the best candidate SNOMED CT concepts. For each algorithm we always began with the original terms, then the UMLS normalized terms, followed by the stemmed terms. During each cycle, we would review the candidate concepts found to determine if it was a match, and if so, what type of match it was based on the algorithm applied. When no matching concepts were found, we would label the term as unmatched. Our experience with the matching algorithms had been that, the sooner we could find a match in the cycle, the greater confidence we would have that the candidate concept is appropriate. The preferred order of matching selected is always exact first, then all, followed by partial. For exact-match and match-all if only inactive concepts are found then a semantic-match is done to find their corresponding current concepts through the historical relationships.

Step-5	Explanation
Stop words	Frequent short words that do not affect the phrase: and, by, for, in, of, on, the, to, with, no, (nos)
Exclude words	Words that may change meaning of the word but if ignored help to find a term otherwise missed: about, alongside, an, anything, around, as, at, because, before, being, both, cannot, chronically, consists, covered, does, during, every, find, from, instead, into, more, must, no, not, only, or, properly, side, sided, some, something, specific, than, that, things, this, throughout, up, using, usually, when, while
SNOMED Prefixes	[X] – concepts with ICD-10 codes not in ICD-9 [D] - concepts in ICD-9 XVI and ICD-10 SVII [M] – morphology of neoplasm concepts in ICD-O [SO] – concepts in OPCS-4 chapter Z in CTV3 [Q] – temporary qualifying terms from CTV3 [V] – concepts in ICD-9 and ICD-10 on factors influencing health status and contact with health services (V-codes and Z-codes)

Table 2b. Expanded UMLS normalization step-2

Encoding the Problem Lists
The process of encoding the problem lists extracted from the EMR followed these steps: (a) tabulating the frequency of occurrences for all of the original PL

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

terms; (b) cataloguing all of the unique words across the PL terms present; (c) examining all unique words and PL terms to identify and revise for acronyms, abbreviations, spelling variants and errors; (d) matching the PL terms to SNOMED CT concepts using matching algorithms described earlier; (e) producing detailed and summary outputs to show the type of matches found; (f) reviewing/verifying the matched concepts one term at a time for accuracy; (g) repeating steps (c-f) until no further matches could be found; (h) examine remaining partial-matches for post-coordination; (i) create an index table of all PL and matched SNOMED terms. As part of this study, we also explored navigating within the SNOMED hierarchy to examine how the super-types and relations could be used to improve the quality of recall using the matched SNOMED concepts.

RESULTS

Summary of PL Terms and Matches

A total of 7,833 PL entries were extracted from the EMR for this study. The majority of these entries were recorded by one practitioner over a 7-year period. Of these entries, there were 1,713 unique PL terms present. Based on the frequency distribution of the entries, the top 10 PL terms were hypertension, hypercholesterolemia, diabetes mellitus, hypothyroid, asthma, atrial fibrillation, gastroesophageal reflux, depression, congestive heart failure and chronic kidney disease. After the second cycle we had 1,296 (88.23%) exact-matches where the PL terms are exactly the same as the SNOMED terms found. There were 51 (3.47%) match-all where all the words in the PL terms are present in the SNOMED terms but not necessarily in the same sequence. There were 120 (8.17%) partial-matches where one or more words matched the SNOMED terms. Another 20 (1.42%) SNOMED terms were found with semantic matches. Between the two cycles partially-matched terms were revised to tease out qualifiers and secondary concepts if present in order to explore post-coordination. A summary of the PL terms and the SNOMED matches found is shown in Table 3.

Characteristics of Encoded PL Terms

In Table 4 we have examples of the frequently used PL terms with their SNOMED terms found by exact, all and semantic matches. Also shown are the matches after revision and post-coordination of the original and partially-matched PL terms. For most exact-matches we selected the preferred terms from SNOMED CT as they are identical or closest to the original PL terms, such as Atrial fibrillation. In some cases we chose the synonym terms, such as Hypertension instead of the preferred term which is Hypertensive disorder. For match-all and some

partial-matches we selected the SNOMED terms that were closest to the PL concept involved, such as GERD gastro-esophageal reflux disease. For semantic matches we looked up the current concepts of the matched but inactive SNOMED terms through their historical relationships, such as Cirrhosis. For post-coordination we added qualifier and refinement terms to SNOMED concepts or combined those that are lexically closest to the original PL terms, such as Atrial fibrillation+Chronic, Kidney disease+Chronic, and Headache+Migraine. After the second cycle any remaining partial-matches were treated as unmatched. Initially there were eight PL terms not found in SNOMED CT. Five were spelling errors and were revised for the second cycle (e.g. hepatomegal*y* → hepatomegaly); three were legitimate missing terms – vasculopath, pyocystitis and hypotestosteronemia, where we had to modify the PL term or tag as local extensions. Using these outputs we created an index table to link the PL terms to their matched SNOMED terms, shown in Table 5. Each row contains the PL-termId, conceptId, descriptionId, relationship-typeId match-type, and post-coordination-sequenceId.

Description	Frequency	
No. of patients	2,894	
Total PL entries	7,833	
Total words in PL terms	16,455	
Unique words	1,764	
Longest word	Hypercholesterolemia, 20 characters	
Median length	8 characters	
Most common word	Hypertension, 585 times	
Matching Algorithm	**Initial Cycle Frequency (%)**	**2nd Cycle Frequency (%)**
Exact-Match	905 (52.83%)	1,296 (88.23%)
Match-All	167 (9.75%)	52 (3.47%)
Partial-Match	633 (36.95%)	120 (8.17%)
Semantic-Match	49 (2.86%)	20 (1.42%)
Unmatched	8 (0.47%)	2 (0.14%)
Post-coordination	Not done	In-progress
Total unique PL terms	**1,713**	**1,468**

Table 3. Summary of PL terms and matches. For frequency %, once a match has been found it is not included as part of the next matching algorithm

Revision of PL Terms

Manual revisions were done on the 1,713 unique PL terms after the initial cycle. By selecting the PL terms that were not matched in SNOMED CT, we were able to identify entries that were misspelled, idiosyncratic local terms or ambiguous concepts. A number of spelling mistakes were corrected. The CliniClue Browser [17] was used to find matches for each term. A few terms were found in our problem lists but not in SNOMED CT. Some were local terms that needed to be reconsidered but there were also terms that would be submitted for inclusion in SNOMED CT. One example is "chronic kidney

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

disease" which seems to be the preferred term in common usage. Yet the closest SNOMED term is "chronic renal failure." In this revision we also noted parts of some PL terms could be removed as qualifiers or modifiers, thus increasing the number of exact matches found. Examples include left, right, lower, midline, chronic, recurring, active, query and multiple. These modifiers seemed to be clustered around the concepts of time course, number, location and severity. We found 313 such instances in our PL terms. In another 89 instances we found post-coordination of two SNOMED concepts produced a good match.

Navigating the SNOMED Hierarchy
As part of this study, we explored ways to navigate the SNOMED hierarchy to determine if it could improve one's ability to retrieve related concepts. Of the 1,296 exact matches found for the 1,468 unique PL terms present, we selected a subset of 32 PL terms related to cardiovascular disorders for this analysis. First, we did frequency counts of these PL terms to show how often they were present in the EMR system. For each PL term present, we navigated up the hierarchy until we reached the super-type "49601007|Disorder of cardiovascular system." We then pruned the tree to include only those concepts with a positive frequency count, but left their immediate super-types intact. This partly-instantiated cardiovascular disorder hierarchy is shown in Figure 1. The value of this tree is that it shows the SNOMED concepts that are actually present in the EMR and how often they occur via the frequency counts based on the PL terms recorded. This tree can aide in the retrieval of relevant concepts recorded using different PL terms. For instance, by specifying the concept "56265001|Heart disease" in the query, one should expect to retrieve all sub-types under "5754005|Acute myocardial infarction" and "12026006|Paroxysmal tachycardia." On the other hand, by specifying the concept "57054005|Acute myocardial infarction" in the query, the sibling concept "12026006|Paroxysmal tachycardia" should automatically be excluded.

<div align="center">

DISCUSSION

</div>

A proposed Methodology
Drawing on the lessons learned from this study, we propose the following steps for general practitioners to encode problem lists from their EMR in SNOMED as a first step for review before full-scale conversion:

1. Extract all PL entries from the EMR and tabulate the frequency of the PL terms present;
2. Catalogue all unique words across the PL terms;

3. Examine all unique words and PL terms to identify and revise for acronyms, abbreviations, spelling variants and errors;
4. Match the PL terms to SNOMED concepts using the matching algorithms outlined in this paper (contact authors for copies of the algorithms);
5. Create detailed and summary outputs to show the exact, all, partial and semantic matches found;
6. Review matched SNOMED terms for accuracy; remove successful exact-match and match-all terms from further matching cycles;
7. Repeat steps 3 through 6 for remaining partial matches until no further matches found;
8. Post-coordinate remaining PL terms with qualifier, refinement and combined concepts;
9. Create a pruned PL hierarchy tree showing all concepts with positive frequency counts and immediate super-type concepts;
10. Create index table containing unique identifiers for the PL and matched SNOMED terms.

Implications
Post-coordination is thought to be a feature that is difficult to implement. Yet based on the small number of SNOMED concepts used in this study to post-coordinate our PL terms, it seems feasible to achieve. We did note the use of pre-coordination in SNOMED CT is unpredictable, and it seems common to include acronyms within SNOMED descriptions. Careful use of modifiers such as laterality, chronicity and severity should be considered. Further studies are needed.

Critics often balk at the unwieldy size and complexity of SNOMED CT as too impractical for local use. In Canada the vendor and general practice communities, which are often small in size, are reluctant to adopt SNOMED CT, questioning their return on value for the effort required. From this study, we have shown it is feasible to incorporate SNOMED CT into EMR in the general practice setting. The methodology we have outlined is practical even for small medical offices with an EMR in place. We have also shown the potential use of SNOMED CT to improve the quality of recall from its hierarchy. The ability to demonstrate return on value, as in our encoding of problem lists with SNOMED CT to improve recall, is an important first step for practitioners to consider before full-scale conversion of their EMR.

Limitations
There are several limitations to this study. First, the PL terms used have been established over the years mainly by one practitioner from a single setting, which are likely to vary between practices. Second, our current matching algorithms do not take into

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

account subtype hierarchy to limit searches, which could otherwise restrict unlikely choices such as Physical Object and Substance. Third, the evaluation of this methodology is incomplete to date; the full extent of the post-coordination effort required to encode the entire set of PL terms in this EMR should be further examined and reported. Fourth, the use of our partly instantiated hierarchy tree to improve recall quality, while promising, requires more thorough investigation into its utility with more complex real-life cases. Its design should also be aligned with the existing SNOMED navigation hierarchy feature that is already in place as part of the new RefSet release.

Next Steps
We are developing a Web-based mapping tool made up of the matching algorithms described earlier to allow the matching of clinical terms to SNOMED CT in an interactive or batch mode. With our focus continued to be on general practice EMR systems, there are several steps ahead to be considered. For instance, we need to expand the use of SNOMED terms to other parts of the EMR such as procedures, medications and billing. We also need to refine our encoding methodology to take into account specific contexts such as past/family history and health risks, and to use subtype hierarchy to improve search precision. The inclusion of frequency statistics on the distribution of matched SNOMED CT terms across the hierarchies would be useful to validate the results. These efforts should aid in the eventual creation of a primary care SNOMED subset, and eventually a concept model in the primary care domain. But most important, we should continue to exploit ways by which the use of SNOMED CT in the EMR can actually enhance patient care.

Original PL Term	Type of Match	Identifier	Id Type	SNOMED Term	Descn Type	Descn Status
Atrial Fibrillation	Exact	49436004	C	Atrial fibrillation (disorder)	F	0
		82343012	D	**Atrial fibrillation**	**P**	**0**
Hypertension	Exact	38341003	C	Hypertensive disorder, systemic arterial (disorder)	F	0
		1215744012	D	Hypertensive disorder	P	0
		64176011	D	**Hypertension**	**S**	**0**
Gastroesophageal Reflux - GERD	All	235595009	C	Gastroesophageal reflux disease (disorder)	F	0
		2535970019	D	**GERD – Gastro-esophageal reflux disease**	**S**	**0**
Cirrhosis	Semantic	155809006	C	Cirrhosis	U	4
		19943007	C	Cirrhosis of liver (disorder)	F	0
		33568015	D	**Cirrhosis of liver**	**P**	**0**
Atrial Fibrillation - Chronic	Post, Exact	**82343012**	D	**Atrial fibrillation**	**P**	**0**
		288524001	C	Courses (qualifier value)	F	0
		428182017	D	**Courses**	**P**	**0**
		90734009	C	Chronic (qualifier value)	F	0
		150360019	D	**Chronic**	**P**	**0**
Chronic Kidney Disease - CKD	Post	90708001	C	Kidney disease (disorder)	F	0
		150315015	D	**Kidney disease**	**P**	**0**
		263502005	C	Clinical course (attribute)	F	0
		391753013	D	**Clinical course**	**P**	**0**
		90734009	C	Chronic (qualifier value)	F	0
		150360019	D	**Chronic**	**P**	**0**
Headache Migraine	Post, Exact	37796009	C	Migraine (disorder)	F	0
		63055014	D	**Migraine**	**P**	**0**
		246090004	C	Associated finding (attribute)	F	0
		367802015	D	**Associated finding**	**P**	**0**
		25064002	C	Headache (finding)	F	0
		41990019	D	**Headache**	**P**	**0**

Table 4. Examples of matched PL and SNOMED terms by exact, all, semantic and post-coordinated matches. Legend: Identifier (contains ConceptId or DescriptionID depending on Id-Type); Id Type (C- Concept, D-Description); Descn-Type (P-preferred, S-synonym, F-fully specified name, U-undefined); Descn-Status (0-current, 4-ambiguous); note that all selected SNOMED terms are shaded and in bold

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

Rec	PL-Id	PL-Term	ConceptId	DescriptionId	Match	AttributeId	SequenceId
1	160	Atrial Fibrillation	49436004	83243013	Exact	0	0
2	789	Hypertension	38341003	64176011	Exact	0	0
3	685	Gastroesophageal Reflux - GERD	235595009	2535970019	All	0	0
4	32666	Chronic Kidney Disease CKD	90708001	150315015	Post	0	0
5	32666	Chronic Kidney Disease CKD	90734009	150360019	Post	263502005	1
6	431	Cirrhosis	19943007	33568015	Semantic	0	0
7	1044	Headache Migraine	37796009	63055014	Post, Exact	0	0
8	1044	Headache Migraine	25064002	41990019	Post, Exact	246090004	1

Table 5. Examples of the index table linking the original PL terms to matched SNOMED terms.
Legend: SequenceId indicates the relative ordering of the post-coordinated records

Two sets of post-coordinated terms shown above

```
49601007  Disorder of cardiovascular system (disorder) - 1
              128487001   Acute disease of cardiovascular system (disorder)
                  127337006   Acute heart disease (disorder)
                      57054005    Acute myocardial infarction (disorder)
                          70211005     Acute myocardial infarction of anterolateral wall (disorder) - 1
                          73795002     Acute myocardial infarction of inferior wall (disorder) - 5
                          307140009    Acute non-Q wave infarction (disorder) - 5
                      12026006   Paroxysmal tachycardia (disorder) - 1
              9904008     Congenital anomaly of cardiovascular system (disorder)
                  363028003   Congenital anomaly of cardiovascular structure of trunk (disorder)
                      13213009    Congenital heart disease (disorder) - 1
                          10818008     Congenital malposition of heart (disorder)
                          27637000  Dextrocardia (disorder) - 1
              27550009    Disorder of blood vessel (disorder)
                  359557001   Disorder of artery (disorder)
                      72092001    Arteriosclerotic vascular disease (disorder)
                          53741008     Coronary arteriosclerosis (disorder) - 9
                      414024009   Disorder of coronary artery (disorder)
                          53741008     Coronary arteriosclerosis (disorder) - 9
              55855009    Disorder of pericardium (disorder)
                      3238004     Pericarditis (disorder) - 2
                      15555002  Acute pericarditis (disorder) - 1
          56265001   Heart disease (disorder) - 1
              127337006   Acute heart disease (disorder)
                  57054005    Acute myocardial infarction (disorder)
                      70211005     Acute myocardial infarction of anterolateral wall (disorder) - 1
                      73795002     Acute myocardial infarction of inferior wall (disorder) - 5
                      307140009    Acute non-Q wave infarction (disorder) - 5
              12026006  Paroxysmal tachycardia (disorder) - 1
```

PL-Id	Original PL Term	Concept Id	Fully Specified Name
4435	Dextrocardia	27637000	Dextrocardia (disorder)
10086	Heart Disease	56265001	Heart disease (disorder)
10087	Heart Disease Congenital	13213009	Congenital heart disease (disorder)
1035	MI Inferior Myocardial Infarction	73795002	Acute myocardial infarction of inferior wall (disorder)
12653	Myocardial Infarction Anterolateral	70211005	Acute anterolateral myocardial infarction (disorder)
1591	Myocardial Infarction Subendocardial (Non Q wave)	307140009	Acute non-Q wave infarction (disorder)
1202	Pericarditis	3238004	Pericarditis (disorder)
13641	Pericarditis Acute	15555002	Acute pericarditis (disorder)
15976	Tachycardia Paroxysmal	12026006	Tachycardia paroxysmal (disorder)

Figure 1. A partial SNOMED hierarchy for cardiovascular disorders derived from a set of original PL terms. The upper figure portion shows the partial SNOMED hierarchy for cardiovascular disorders; the lower figure portion shows the original PL terms with the matched SNOMED concepts and their fully specified names. In the hierarchy, concepts that are bold and italicized are exact matches for the PL terms, followed by the frequency of how often they appeared in the EMR.

Representing and sharing knowledge using SNOMED
Proceedings of the 3rd international conference on Knowledge Representation in Medicine (KR-MED 2008)
R. Cornet, K.A. Spackman (Eds)

REFERENCES

1. Chute CG, Elkin PL, Fenton SH, Atkin GE. A clinical terminology in the post modern era: pragmatic problem list development. *Proceedings AMIA Ann Symposium* 1998; 795-9.

2. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. *Proceedings AMIA Annual Symposium* 1998; 280-4.

3. Petersson H, Gunnar N, Strender LE, Ahlfeldt H. The connection between terms used in medical records and coding system: a study on Swedish primary health care data. *Medical Informatics* 2001; 26(2):87-99.

4. Wang SJ, Bates DW, Chueh HC, Karson AS, Maviglia SM, Greim JA, Frost JP, Kuperman GJ. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. *International Journal of Medical Informatics* 2003; 72, 17-28.

5. Fabry P, Baud R, Ruch P, Despont-Gros C, Lovis C. Methodology to ease the construction of a terminology of problems. *International Journal of Medical Informatics* 2006;75:624-32.

6. Meystre S, Haug PJ. Automation of a problem list using natural language processing. *BMC Medical Informatics and Decision Making* 2005;5(30):1472-6947/5/30.

7. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clinic Proceedings* 2006;81(6):741-8.

8. O'Halloran J, Miller GC, Britt H. Defining chronic conditions for primary care with ICPC-2. *Family Practice* 2004;21(4):381-6.

9. Wasserman H, Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *Proceedings AMIA Symposium* 2003, 699-703.

10. Comment: one author reviewed datasets provided by colleagues from hospitals in Buenos Aires in Argentina, Kaiser Permanente in United States, and Sherbrooke in Canada during Fall 2007.

11. Comment: WONCA is the World Organization of Family Doctors, and WICC is the WONCA International Classification Committee. URL http://www.globalfamilydoctor.com/; Jan20/08.

12. Lee DHK. *Reverse Mapping ICD-10-CA to SNOMED CT*. UVic Master of Science research project report, Oct 2007. Unpublished.

13. Wang Y, Patrick J, Miller G, O'Halleran. Linguistic mapping of terminologies to SNOMED CT. *Semantic Mining Conference on SNOMED CT* Oct 2006, Copenhagen, Denmark.

14. Kleinsorge R, Willis J, et al. UMLS Overview – Tutorial T12. *AMIA Annual Symposium* 2006. http://165.112.6.70/research/umls/pdf/AMIA_T12_2006_UMLS.pdf. Jan15/08.

15. National Library of Medicine. *The SPECIALIST Lexicon*. http://lexsr3.nlm.nih.gov/LexSysGroup/Projects/Summary/lexicon.html. Jan15/2008.

16. Goldsmith JA, Higgins D, Soglasnova S. *Automatic Language-specific Stemming in Information Retrieval.* Springer-Verlag Berlin Heidelberg 2001.

17. CliniClue. The Clinical Information Consultancy, Ltd., UK. http://www.cliniclue.com/software. available for download. Jan22/08.